# Combined result and associated uncertainty from interlaboratory evaluations based on the ISO *Guide*

*R. Kacker, R. Datla and A. Parr*

**Abstract.** We address the problem of determining the combined result and its associated uncertainty in the measurement of a common measurand by a group of competent laboratories. Most data analyses of interlaboratory evaluations are based on the assumption that the expected values of the individual laboratory results are all equal to the value of the common measurand. This means that the laboratory results are subject to random effects only with respect to the value of the measurand. This assumption is frequently unjustified. We use the more realistic assumption that the laboratory results are subject to both random and systematic effects with respect to the value of the measurand. In this case, the value of the measurand may fall anywhere within the range of results. Therefore, a combined result and its associated standard uncertainty that place a non-negligible fraction of the results outside the 2-standard-uncertainty interval are unsatisfactory representations of the value of the common measurand provided by the set of laboratory results. The more realistic assumption requires us to deal with the uncertainty arising from possible systematic effects in the laboratory results. Following the approach of the ISO *Guide* to deal with systematic effects, we propose a three-step method to determine a combined result and its associated standard uncertainty such that the 2-standard-uncertainty interval would include a sufficiently large fraction of the results. When the interlaboratory evaluation is an International Committee for Weights and Measures (CIPM) key comparison, we suggest that the combined result and its associated standard uncertainty determined by the three-step method be identified with the key comparison reference value and its associated standard uncertainty. These quantities can then be used to specify the degree of equivalence of the individual laboratory results. We illustrate the three-step method by applying it to the results of an international comparison of cryogenic radiometers recently organized by the Consultative Committee for Photometry and Radiometry (CCPR).

## 1. Introduction

We address the objective of determining the combined result and its associated standard uncertainty in the measurement of a common measurand by a group of competent laboratories. We assume that all laboratory results denoted by $x_1$, ..., $x_n$ are valid. Any suspect results have been withdrawn or revised in accordance with the protocol of the interlaboratory evaluation. As far as possible, recognized effects of different experimental conditions in the participating laboratories and recognized changes in the value of the common measurand, $Y$, have been accounted for in results $x_1$, ..., $x_n$ and their associated standard uncertainties $u(x_1)$, ..., $u(x_n)$. The combined result, $y$, and its associated standard uncertainty, $u(y)$, represent the information about $Y$ provided by $x_1$, ..., $x_n$ and $u(x_1)$, ..., $u(x_n)$.

We follow the philosophy, terminology and notation of the *Guide to the Expression of Uncertainty in Measurement*, referred to here as the ISO *Guide* [1], published by the International Organization for Standardization (ISO) and supported by seven international scientific organizations. The following terms are defined in Appendix 1 of this paper: *measurand, result of a measurement, uncertainty of measurement, standard uncertainty, combined standard uncertainty, expanded uncertainty, coverage factor, expanded uncertainty interval*, and *measurement equation*. We use a coverage factor of two ($k = 2$) and refer to the expanded uncertainty interval as a 2-standard-uncertainty interval. When the distribution represented by a result and its associated standard uncertainty is taken to be normal (Gaussian), the 2-standard-uncertainty interval represents about 95 % level of confidence or coverage probability.

At a meeting held in Paris on 14 October 1999, the directors of the national metrology institutes (NMIs) of thirty-eight Member States of the Metre Convention and representatives of two international organizations signed a Mutual Recognition Arrangement (MRA) [2]. The objectives of the MRA are as follows:

- to establish the degree of equivalence of national measurement standards maintained by NMIs;

R. Kacker, R. Datla and A. Parr: National Institute of Standards and Technology (NIST), Gaithersburg, MD 20899, USA.

- to provide for the mutual recognition of calibration and measurement certificates issued by NMIs;
- thereby to provide governments and other parties with a secure technical foundation for wider agreements related to international trade, commerce and regulatory affairs.

Key comparisons are special international interlaboratory evaluations that serve as the technical basis of the MRA. Key comparisons carried out by the Consultative Committees of the CIPM or the Bureau International des Poids et Mesures (BIPM) are referred to as CIPM key comparisons. The outputs of a CIPM key comparison are as follows: a key comparison reference value (KCRV) (assigned to the common measurand) accompanied by its uncertainty and the degrees of equivalence of all individual measurement standards (laboratory results) accompanied by their uncertainties.

Most data analyses of interlaboratory evaluations, including CIPM key comparisons, are based on the following assumption.

*Assumption I*: The expected values of results $x_1$, ..., $x_n$ are all equal. The common expected value is equal to the value $Y$ of the common measurand. This means that results $x_1$, ..., $x_n$ are subject to random effects only with respect to the value $Y$ of the measurand.

Random effects arise from unpredictable or stochastic temporal and spatial variations of influence quantities. An influence quantity affects the result of the measurement but is not the measurand. Independent measurements reduce the uncertainty arising from random effects. Systematic effects are effects of influence quantities that give rise to uncertainty that cannot be reduced by more measurements. Assumption I is frequently unjustified because it presumes that results $x_1$, ..., $x_n$ are completely free of systematic effects. We use the following more realistic assumption.

*Assumption II*: The expected values of results $x_1$, ..., $x_n$ may not all equal the value $Y$ of the common measurand. This means that results $x_1$, ..., $x_n$ are subject to both random and systematic effects with respect to the value $Y$ of the measurand. It is believed, based on scientific judgement about the method(s) of measurement, that the systematic effects are such that the value $Y$ of the measurand is either somewhere in the range of results $x_1$, ..., $x_n$ or in the vicinity of this range when $n$ is small.

According to Assumption I, the unknown value $Y$ of the measurand is believed to be close to the "best combined result" determined by statistical analysis. The results that end up outside the 2-standard-uncertainty interval associated with the combined result are believed to deviate because of random error. Therefore, the fraction of the results excluded by the 2-standard-uncertainty interval is of no consequence.

According to Assumption II, the unknown value $Y$ of the measurand may be anywhere in the range of results $x_1$, ..., $x_n$, or even outside this range when $n$ is small. Therefore, a combined result and its associated standard uncertainty that place a non-negligible fraction of the results outside the 2-standard-uncertainty interval are unsatisfactory representations of the information about $Y$ provided by $x_1$, ..., $x_n$.

Let us consider an example. Figure 1 displays a subset of the results from a recent interlaboratory evaluation involving seventeen national metrology institutes. Let us suppose, for illustration, that the eighteen circles in Figure 1 are the results of direct measurements of a common measurand of value $Y$ by the seventeen laboratories and the arithmetic mean of these results. (Section 5 shows that the results are not direct measurements of a common measurand.) The arithmetic mean of the seventeen results is the uncorrected combined result (u.c.r.). The associated 2-standard-uncertainty intervals are shown as vertical bars. The centre line is drawn at the arithmetic mean. The dashed lines are drawn at the limits of the 2-standard-uncertainty interval associated with the arithmetic mean. The seventeen results (circles) are the best estimates of the common value $Y$ of the measurand provided by highly competent laboratories. We note that the 2-standard-uncertainty interval $[x_\mathrm{A} \pm 2u(x_\mathrm{A})]$ associated with the arithmetic mean $x_\mathrm{A}$ excludes five (i.e. 29 %) of the seventeen results $x_1$, ..., $x_{17}$. Also, several other results are on the borderline. Similarly, the 2-standard-uncertainty interval associated with the weighted mean, with weights proportional to the reciprocals of the squared standard uncertainties, clearly excludes eight (i.e. 47 %) of the seventeen results. According to Assumption II, the unknown

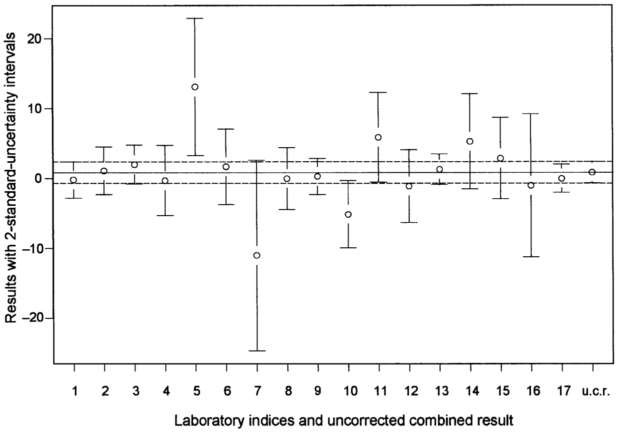

**Figure 1.** Individual laboratory results and uncorrected combined result $x_\mathrm{A}$ with associated 2-standard-uncertainty intervals. Table 1 lists the laboratories corresponding to indices 1, ..., 17 and u.c.r. stands for uncorrected combined result. The horizontal centre line is drawn at the uncorrected combined result $x_\mathrm{A}$. The dashed lines are drawn at the limits $x_\mathrm{A} - 2u(x_\mathrm{A})$ and $x_\mathrm{A} + 2u(x_\mathrm{A})$ of the 2-standard-uncertainty interval associated with $x_\mathrm{A}$.

value $Y$ may be anywhere in the range of the seventeen results. Therefore, the 2-standard-uncertainty intervals associated with the arithmetic mean or the weighted mean are unsatisfactory representations of the information about $Y$ provided by the seventeen results and their associated uncertainties.

Assumption II requires us to deal with the uncertainty about the value $Y$ of the common measurand arising from possible systematic effects in results $x_1, ..., x_n$. Following the ISO *Guide* on dealing with systematic effects, we propose a three-step method to determine a combined result $y$ and its associated standard uncertainty $u(y)$ such that the 2-standard-uncertainty interval $[y \pm 2u(y)]$ would include a sufficiently large fraction of the results. In the case of a CIPM key comparison, we suggest that the combined result and its associated standard uncertainty be identified with the KCRV and its associated standard uncertainty.

In Section 2, we discuss the relevant literature and give an outline of the proposed three-step method. Section 3 gives details of the three-step method and Section 4 discusses how the combined result and its associated standard uncertainty determined by the three-step method can be used to specify the KCRV and the degree of equivalence in CIPM key comparisons. In Section 5, we illustrate the three-step method by applying it to the results of an international comparison of cryogenic radiometers recently organized by the CCPR. A summary is given in Section 6.

## 2. Literature review and proposed three-step method

The early literature on interlaboratory evaluations includes Birge [3] in *Physical Review* and Cochran [4] in *Journal of the Royal Statistical Society*. Since Birge and Cochran, most of the literature on interlaboratory evaluations deals with random effects only. We propose an approach that addresses both random and systematic effects in the laboratory results. Two previous publications that account for the uncertainty arising from possible systematic effects are Schiller and Eberhardt [5] and Levenson et al. [6]. Schiller and Eberhardt modified the method of Paule and Mandel [7]. These articles were primarily written to assign a *certified value* and its associated uncertainty to a *standard reference material* (SRM) from multiple methods of measurement. The problem they address is similar to our problem where the measurement methods correspond to the laboratories.

Paule and Mandel address the case where the laboratory results $x_1, ..., x_n$ are arithmetic means of independent measurements. The number of measurements made in each laboratory may differ. They assume that the laboratory results $x_1, ..., x_n$ are independent random variables with the same expected value that is equal to the unknown value of the common measurand but with different standard deviations. This accords with Assumption I. They model the variances (squared standard deviations) of the laboratory results as consisting of two parts: a common between-laboratory variance and different within-laboratory variances. They estimate the within-laboratory variances as the experimental (sample) variances of the arithmetic means $x_1, ..., x_n$. Then they estimate the between-laboratory variance as that quantity which makes the square of the Birge ratio equal to one. (We describe the Birge ratio in Section 3.) Such estimate of between-laboratory variance is determined by iteration, starting with a guesstimate. The weights in the Birge ratio are iterative estimates of the reciprocals of the variances of $x_1, ..., x_n$. In most cases only a few iterations are required. The final estimate of the between-laboratory variance and the estimates of the within-laboratory variances give the estimated variances of $x_1, ..., x_n$. Paule and Mandel recommend a weighted mean of the laboratory results $x_1, ..., x_n$ as the combined result denoted by $x_{\mathrm{C}}$. The weights are proportional to the reciprocals of the estimated variances of $x_1, ..., x_n$. For determining the standard uncertainty associated with the combined result $x_{\mathrm{C}}$, they treat the weights used in $x_{\mathrm{C}}$ and the estimated variances of results $x_1, ..., x_n$ as constants.

Rukhin and Vangel [8] show that the Paule-Mandel result may be interpreted approximately as the *maximum likelihood estimate* based on the *one-way random effects analysis of variance* (ANOVA) model with normally distributed within-laboratory and between-laboratory effects, unequal numbers of measurements, and different within-laboratory standard deviations. Thus, Rukhin and Vangel put Paule-Mandel on a solid statistical foundation. An expanded uncertainty interval determined by the method of Paule and Mandel may exclude a non-negligible fraction of results $x_1, ..., x_n$.

Motivated by the desire to include all laboratory results, Schiller and Eberhardt [5] modified the expanded uncertainty interval associated with the Paule-Mandel weighted mean $x_{\mathrm{C}}$. They estimate the expanded uncertainty arising from random error in the Paule-Mandel weighted mean $x_{\mathrm{C}}$ from the estimates of within-laboratory standard deviations. Then they estimate the extent of possible systematic error in the Paule-Mandel weighted mean $x_{\mathrm{C}}$ and refer to it as the bias allowance. They set the bias allowance as the maximum absolute deviation $max\{|x_1 - x_{\mathrm{C}}|, \ |x_2 - x_{\mathrm{C}}|, ..., |x_n - x_{\mathrm{C}}|\}$ of any laboratory result $x_1, ..., x_n$ from $x_{\mathrm{C}}$. The bias allowance recognizes that the unknown value $Y$ may be close to any one of results $x_1, ..., x_n$. Schiller and Eberhardt add the bias allowance to the expanded uncertainty arising from random error. The expanded uncertainty interval so obtained is sufficiently wide to include all laboratory results. This approach and its unpublished modifications have often been used to assign a certified value and its associated expanded uncertainty to SRMs at the National Institute of Standards and Technology (NIST). However, the

Schiller-Eberhardt procedure is not consistent with the ISO *Guide* and NIST TN-1297 [9].

Motivated by the desire to be consistent with the ISO *Guide*, Levenson et al. [6] address the case of two laboratories where the arithmetic mean of the two results is used as the combined result. (They refer to it as a two-method problem and suggest that their solution can be used even when the number of methods is three or four. However, they do not give details for doing this.) This method is often used at the NIST to assign a certified value and its associated uncertainty to the SRMs from two independent methods. Levenson et al. use the arithmetic mean $x_{\mathrm{A}} = (x_1 + x_2)/2$ of the two results $x_1$ and $x_2$ as the combined result. They denote the expected value of the combined result $E(x_{\mathrm{A}}) = [E(x_1) + E(x_2)]/2$ by $\mu$, and define the standard uncertainty $u(x_{\mathrm{A}})$ associated with $x_{\mathrm{A}}$ as a statistical estimate of the standard deviation $S(x_{\mathrm{A}})$. The uncertainty $u(x_{\mathrm{A}})$ represents the uncertainty arising from random effects. Levenson et al. model the value of the common measurand denoted by $\gamma$ as $\gamma = \mu + \beta$, where $\beta$ is the bias (systematic error) in $x_{\mathrm{A}}$. The existence of bias $\beta$ means that at least one of the expected values $E(x_1)$ and $E(x_2)$ is not equal to the value of the measurand. This accords with Assumption II. In order to determine an estimate of the common measurand $\gamma$ and the associated standard uncertainty, Levenson et al. assume that $\beta$ is a random variable with rectangular or normal distribution having the expected value $E(\beta) = 0$ and the standard deviation $S(\beta)$. They use the spread of the results $x_1$ and $x_2$ to quantify $S(\beta)$. As the expected value $E(\beta)$ is assumed to be zero, the arithmetic mean $x_{\mathrm{A}}$ is an estimate of both $\mu$ and $\gamma$. The uncertainty associated with $x_{\mathrm{A}}$ treated as an estimate of $\gamma$ has two components: $u(x_{\mathrm{A}})$ and $S(\beta)$, which are combined by the root-sum-of-squares method. Thus the Levenson et al. estimate of the common measurand $\gamma$ is the arithmetic mean $x_{\mathrm{A}}$ with standard uncertainty $\sqrt{[u^2(x_{\mathrm{A}}) + S^2(\beta)]}$. The corresponding expanded uncertainty interval includes both $x_1$ and $x_2$. Levenson et al. seem to treat $\mu = E(x_{\mathrm{A}})$ as an unknown constant. Their model $\gamma = \mu + \beta$ and the assumption that $\beta$ is a random variable with expected value $E(\beta) = 0$ and standard deviation $S(\beta)$ imply that $\gamma$ is a random variable with expected value $E(\gamma) = \mu$ and standard deviation $S(\gamma) = S(\beta)$. However, Levenson et al. seem to conclude that $S(\gamma) = \sqrt{[u^2(x_{\mathrm{A}}) + S^2(\beta)]}$. We note that their conclusion would follow immediately from their model $\gamma = \mu + \beta$ if they treated not only $\beta$ but also $\mu$ as an independent random variable with expected value $x_{\mathrm{A}}$ and standard deviation $u(x_{\mathrm{A}})$. Levenson et al. discuss a Bayesian analysis in Appendix B of their paper.

We address the general case where the number $n$ of laboratories is arbitrary and the combined result may be based on the arithmetic mean or a weighted mean. We seek a combined result $y$ for the value $Y$ of the common measurand and the uncertainty $u(y)$ such that the 2-standard-uncertainty interval $[y \pm 2u(y)]$ would include a sufficiently large fraction of results $x_1$, ..., $x_n$. In order to determine such $y$ and $u(y)$, we need to account for the uncertainty arising from possible systematic effects in results $x_1$, ..., $x_n$ with respect to the value $Y$ of the measurand.

Until the publication of the ISO *Guide* and NIST TN-1297, there was no generally accepted approach to account for the uncertainty arising from systematic effects. The ISO *Guide* (Section 3.2) and NIST TN-1297 (Section 5.2) recommend that each result should be corrected for all recognized systematic effects and that every effort should be made to identify such effects. The uncertainty associated with each result should include both the uncertainties associated with the corrections applied for systematic effects and the uncertainty arising from random effects. The correction applied for a systematic effect and the uncertainty associated with the correction are generally determined from a probability distribution that represents belief about reasonable correction. The ISO *Guide* further recommends that all uncertainty components, whether arising from random effects or from corrections for systematic effects, should be expressed as standard deviations. The result of a measurement (including corrections for systematic effects) is determined from a measurement equation. The combined standard uncertainty (including both random and systematic uncertainties) is determined from the law of propagation of uncertainties (also called the root-sum-of-squares method). The law of propagation of uncertainties is derived from a first-order Taylor series approximation of the measurement equation. When the law is believed to be inadequate, the combined standard uncertainty may be determined from a numerical simulation of the measurement equation.

We propose a three-step method based on the ISO *Guide*.

*Step 1*: *Determine the uncorrected combined result $x_{\mathrm{C}}$ and its associated standard uncertainty $u(x_{\mathrm{C}})$. Assess the need for correcting the result $x_{\mathrm{C}}$.* The uncorrected combined result $x_{\mathrm{C}}$ is generally the arithmetic mean or a weighted mean of the individual results $x_1$, ..., $x_n$. Determine the corresponding standard uncertainty $u(x_{\mathrm{C}})$ from the uncertainties $u(x_1)$, ..., $u(x_n)$ associated with results $x_1$, ..., $x_n$ using the law of propagation of uncertainties. (The uncertainty $u(x_{\mathrm{C}})$ should include covariance terms when some of the laboratory values are correlated.) Assess the need for correcting the result $x_{\mathrm{C}}$ for a possible difference between $x_{\mathrm{C}}$ and the value $Y$ of the measurand. Such correction is needed whenever the interval $[x_{\mathrm{C}} \pm 2u(x_{\mathrm{C}})]$ excludes a non-negligible fraction of the results.

*Step 2*: *Determine the correction to be applied to $x_{\mathrm{C}}$ and the standard uncertainty associated with the correction.* Following the ISO *Guide* (Section 4.3), a probability distribution is used to determine the correction to be applied and the uncertainty associated with it. The probability distribution represents belief about reasonable correction. We discuss two simple and useful distributions. The parameters of the probability distribution of correction are determined from the range of the deviations $(x_1 - x_\mathrm{C})$, $(x_2 - x_\mathrm{C})$, ..., $(x_n - x_\mathrm{C})$ of the results $x_1, x_2$, ..., $x_n$ from the uncorrected combined result $x_\mathrm{C}$.

*Step 3*: *Determine the corrected combined result and its associated combined standard uncertainty*. The corrected combined result $y$ is obtained by applying a correction to $x_\mathrm{C}$. The associated standard uncertainty $u(y)$ is determined by combining the uncertainty $u(x_\mathrm{C})$ and the uncertainty associated with the correction using the law of propagation of uncertainties. With a judiciously specified probability distribution for the correction, the 2-standard-uncertainty interval $[y \pm 2u(y)]$ would include a sufficiently large fraction of results $x_1$, ..., $x_n$. Thus $y$ and $u(y)$ would represent, better than $x_\mathrm{C}$ and $u(x_\mathrm{C})$, the information about $Y$ provided by results $x_1$, ..., $x_n$. The uncertainty $u(y)$ about the value $Y$ of the common measurand is greater than the uncertainty $u(x_\mathrm{C})$ because Assumption II presumes less than Assumption I.

We use upper-case symbols, such as $Y$, $X_1$, ..., $X_n$, and $C_{i1}$, $C_{i2}$, ..., for random variables that represent states of knowledge. The symbol $Y$ is used for the unknown value of the measurand as well as the random variable that represents the state of knowledge about the value of the measurand. We use lower-case symbols, such as $y$, $x_1$, ..., $x_n$, $x_\mathrm{C}$, $x_\mathrm{A}$, $x_\mathrm{W}$, $c_{i1}$, $c_{i2}$, ..., and $z_1$, ..., $z_n$, for known quantities, evaluated from statistical data (Type A) and/or scientific judgement (Type B), and for random variables with sampling distributions from the measurement process. Greek symbols, such as $\mu$, $\sigma$, $\alpha$ and $\beta$, may be unknown parameters, set parameters or random variables representing states of knowledge. The context makes clear the way in which a symbol is being used. The statistical functions $E(Z)$, $V(Z)$, and $S(Z)$ are respectively the expected value, the variance, and the standard deviation of the argument denoted here by random variable $Z$.

## 3. Details of three-step method

Frequently, one or more of the results from interlaboratory evaluations are relatively remote from the rest and are referred to as discrepant. A set of laboratory results that contains discrepant results is said to be inconsistent. A classic method of checking the consistency of a set of interlaboratory results is the Birge ratio test [3] (see, for example, Taylor et al. [10] and Mohr and Taylor [11]).

### 3.1 Birge ratio test of consistency

The Birge ratio denoted by $R_\mathrm{B}$ is defined as $R_\mathrm{B} = \sqrt{[\sum_i w_i (x_i - x_\mathrm{W})^2/(n-1)]}$, where $x_1$, ..., $x_n$ are the laboratory results and $x_\mathrm{W}$ is the weighted mean, $\sum_i w_i x_i / \sum_i w_i$, with weights $w_i = 1/u^2(x_i)$ for $i = 1$, ..., $n$. Taylor et al. give a brief and perceptive description of the Birge ratio and its original interpretation. Appendix 2 of this paper describes the Birge ratio as a statistical estimate. Here we discuss its use.

The statistical model underlying the Birge ratio $R_\mathrm{B}$ is that the laboratory results $x_1$, ..., $x_n$ are independently distributed random variables with a common unknown expected value denoted by $E(x_i) = \mu$ but with different known standard deviations $S(x_i) = u(x_i)$ for $i = 1$, ..., $n$. Consistency means that results $x_1$, ..., $x_n$ and standard uncertainties $u(x_1)$, ..., $u(x_n)$ fit the Birge ratio model and inconsistency means that they do not fit the model, i.e. some results are discrepant. It can be shown that when results $x_1$, ..., $x_n$ and standard uncertainties $u(x_1)$, ..., $u(x_n)$ fit the Birge ratio model, $E(R_\mathrm{B}^2) = 1$ (see Appendix 2). Therefore, the values of $R_\mathrm{B}$ that are close to 1 or less suggest that results $x_1$, ..., $x_n$ are consistent. The values of $R_\mathrm{B}$ that are much greater than 1 suggest that results $x_1$, ..., $x_n$ are inconsistent.

A discrepant result may be erroneous or an outlier. An outlier is a discrepant result that has not been determined to be erroneous. All discrepant results should be critically investigated according to the protocol of the interlaboratory evaluation. Results determined to be erroneous should be revised or removed. Often, the resources and time available for investigations are limited and some discrepant results may remain as outliers. The combined result $y$ and the uncertainty $u(y)$ should account for the uncertainty arising from such outliers.

In the Birge ratio test, the standard uncertainties $u(x_1)$, ..., $u(x_n)$ are treated as known parameters representing the standard deviations of laboratory results $x_1$, ..., $x_n$. That is, the Birge ratio test requires an implicit assumption that each of the standard uncertainties $u(x_1)$, ..., $u(x_n)$ is reliable. When this assumption is not well justified, the conclusion of the Birge ratio test should not be taken too seriously. The best statistical estimate of $\mu$ in the Birge ratio model is the weighted mean $x_\mathrm{W} = \sum_i w_i x_i / \sum_i w_i$, where $w_i = 1/u^2(x_i)$ for $i = 1$, ..., $n$. Again, this requires the assumption that each of the uncertainties $u(x_1)$, ..., $u(x_n)$ is reliable.

The Birge ratio consistency does not affirm Assumption I because the common expected value $\mu$ in the Birge ratio model need not be the unknown value $Y$ of the measurand. The first part of Assumption I, that "the expected values of results are all equal" is a statement about consistency. The second part, that "the common expected value is equal to the value $Y$ of the common measurand" is a statement about accuracy. Consistency does not imply accuracy.

### 3.2 Corrected laboratory results $x_1$, ..., $x_n$ and associated standard uncertainties $u(x_1)$, ..., $u(x_n)$

According to the ISO *Guide*, the individual laboratory results $x_1$, ..., $x_n$ should be corrected for all recognized systematic effects and their associated standard uncertainties $u(x_1)$, ..., $u(x_n)$ should include the uncertainties arising from random effects and the uncertainties associated with the corrections. Here is the procedure.

We use lower-case symbols $z_1$, ..., $z_n$ to denote the arithmetic means of the raw measurements denoted by $\{z_{ij}\}$. The laboratory results $x_1$, ..., $x_n$ are obtained from the arithmetic means $z_1$, ..., $z_n$ by applying corrections for recognized systematic effects.

In classical frequentist statistics, the uncertainty from random effects based on a series of $n_i$ independent measurements $\{z_{ij}\}$ in a particular laboratory, $i$, for $i =$ 1, 2, ..., $n$, is determined as follows. The measurements $\{z_{ij}\}$ for fixed $i$ are modelled as $z_{ij} = \mu_i + e_{ij}$, where $E(e_{ij}) = 0$, $V(e_{ij}) = \sigma_i^2$, and $j = 1, ..., n_i$. Then $E(z_{ij}) = \mu_i$ and $V(z_{ij}) = \sigma_i^2$. Let $z_i = (1/n_i)\sum_j z_{ij}$ and $s_i^2 = \sum_j (z_{ij} - z_i)^2/(n_i - 1)$. Then $E(z_i) = \mu_i$, $V(z_i) = \sigma_i^2/n_i$, $S(z_i) = \sigma_i/\sqrt{n_i}$, $E(s_i^2/n_i) = V(z_i)$ (see, for example, Hoel [12]). Thus $z_i$ is an estimate of $\mu_i$ and $s_i/\sqrt{n_i}$ is an estimate of $S(z_i)$. In classical frequentist statistics, one takes the estimate $s_i/\sqrt{n_i}$ of $S(z_i)$ as the standard uncertainty $u(z_i)$ of $z_i$. When the measurements $\{z_{ij}\}$ can be assumed to be normally distributed, the ratio $(z_i - \mu_i)/(s_i/\sqrt{n_i})$ has Student's $t$-distribution with $(n_i - 1)$ degrees of freedom. This result is used to form a confidence interval for $\mu_i$. Here $\mu_i$ is interpreted as an unknown constant and the confidence interval is random.

In Bayesian statistics, one thinks in terms of the state of knowledge – expressed as a probability distribution – about the parameters $\mu_i$ and $\sigma_i$ treated as random variables. One starts with prior distributions for $\mu_i$ and $\sigma_i$ that represent the states of knowledge about them before measurements are taken. The measurements $\{z_{ij}\}$ are used to update the prior distribution to obtain a posterior distribution. The mechanism for updating is Bayes theorem (see, for example, Box and Tiao [13]). Negligible prior knowledge about $\mu_i$ and $\sigma_i$ is expressed by using non-informative prior distributions. When the prior distributions are non-informative and $\{z_{ij}\}$ are normally distributed, the posterior distribution of the ratio $(\mu_i - z_i)/(s_i/\sqrt{n_i})$ turns out to be Student's $t$-distribution with $(n_i - 1)$ degrees of freedom [13]. Here $\mu_i$ is interpreted as a random variable and $z_i$ and $s_i/\sqrt{n_i}$ are known quantities. It follows that $E(\mu_i) = z_i$ and $S(\mu_i) = \sqrt{[(n_i - 1)/(n_i - 3)]} \times (s_i/\sqrt{n_i})$, provided that $n_i > 3$ (see, for example, Evans et al. [14]). Thus in Bayesian statistics, the arithmetic mean $z_i$ is $E(\mu_i)$ and $\sqrt{[(n_i - 1)/(n_i - 3)]} \times (s_i/\sqrt{n_i})$ is $S(\mu_i)$. So in Bayesian statistics, one would take $\sqrt{[(n_i - 1)/(n_i - 3)]} \times (s_i/\sqrt{n_i})$ as the standard uncertainty $u(z_i)$ associated with the arithmetic mean $z_i$. The factor $\sqrt{[(n_i - 1)/(n_i - 3)]}$ built into the Bayesian standard uncertainty accounts for the uncertainty that arises when $n_i$ is small.

The ISO *Guide* does not cite the Bayesian result that, given $z_i$ and $s_i/\sqrt{n_i}$, the ratio $(\mu_i - z_i)/(s_i/\sqrt{n_i})$ has Student's $t$-distribution. However, Section 6.2.2 of the *Guide* mentions the distribution characterized by the result of measurement and its combined standard uncertainty and defines the coverage probability or the level of confidence as the fraction of the distribution covered by the expanded uncertainty interval. This definition corresponds to the Bayesian viewpoint of treating $\mu_i$ as a random variable representing the state of knowledge. We have adopted this Bayesian viewpoint. Thus we interpret the arithmetic mean $z_i$ as the expected value $E(\mu_i)$ with standard uncertainty $u(z_i) = S(\mu_i) = \sqrt{[(n_i - 1)/(n_i - 3)]} \times (s_i/\sqrt{n_i})$ for $i = 1$, 2, ..., $n$. In the rest of this section, the arithmetic means $z_1$, ..., $z_n$, uncertainties $u(z_1)$, ..., $u(z_n)$, laboratory results $x_1$, ..., $x_n$, and uncertainties $u(x_1)$, ..., $u(x_n)$ are known quantities rather than random variables, with one exception. When we discuss the suitability of the weighted mean and the median as a reference value, results $x_1$, ..., $x_n$ are random variables.

A measurement equation is used to incorporate corrections for recognized systematic effects. Let $C_{i1}$, $C_{i2}$, ... be independent corrections for recognized systematic effects in $z_i$. Then the measurement equation for the random variable $X_i$ that represents the state of knowledge about the value $Y$ of the common measurand in laboratory $i$ is $X_i = \mu_i + C_{i1} + C_{i2} + \ldots = \mu_i + \sum_j C_{ij}$, where all terms are random variables representing states of knowledge. Let $E(C_{ij}) = c_{ij}$ and $S(C_{ij}) = u(c_{ij})$ for $i = 1$, 2, ..., $n$ and $j = 1$, 2, ... . The random variables $C_{i1}$, $C_{i2}$, ... and their expected values $c_{i1}$, $c_{i2}$, ... are both referred to as corrections. The corrected result $x_i$ for the laboratory $i$ is $x_i = z_i + c_{i1} + c_{i2} + \ldots = z_i + \sum_j c_{ij}$ with uncertainty $u(x_i) = \sqrt{[u^2(z_i) + \sum_j u^2(c_{ij})]}$. The result $x_i$ includes the corrections $c_{i1}$, $c_{i2}$, ... applied for recognized systematic effects in the arithmetic mean $z_i$ of the raw measurements $\{z_{ij}\}$. The uncertainty $u(x_i)$ includes the uncertainty $u(z_i)$ arising from random effects and the uncertainties $u(c_{i1})$, $u(c_{i2})$, ... associated with the corrections for $i = 1$, 2, ..., $n$.

The random variables $X_1$, ..., $X_n$ represent the states of knowledge in the $n$ laboratories about the value $Y$ of the common measurand. We refer to them as laboratory values of the measurand. The result $x_i$ from laboratory $i$, for $i = 1$, 2, ..., $n$, is identified with the expected value $E(X_i)$ and the uncertainty $u(x_i)$ is identified with the standard deviation $S(X_i)$. Often, uncertainty is expressed as relative standard uncertainty denoted by $u_{\mathrm{r}}(x_i)$, where $u_{\mathrm{r}}(x_i) = u(x_i)/|x_i|$ provided that $x_i$ is not zero for $i = 1$, 2, ..., $n$. The relative standard uncertainty has the advantage of being a dimensionless quantity frequently expressed as a percentage.

### 3.3 Uncorrected combined result $x_{\mathrm{C}}$ and associated standard uncertainty $u(x_{\mathrm{C}})$

We use the symbol $X_{\mathrm{C}}$ to represent a function $X_{\mathrm{C}} = f(X_1, ..., X_n)$ of the laboratory values

$X_1, ..., X_n$ that is used as the uncorrected combined value of the common measurand. The function $X_{\rm C} = f(X_1, ..., X_n)$ is a measurement equation that defines $X_{\rm C}$. Following the ISO *Guide* (Section 4.1), the corresponding uncorrected combined result denoted by $x_{\rm C}$ is defined as $x_{\rm C} = f(x_1, ..., x_n)$. We assume that the laboratory values $X_1, ..., X_n$ are mutually uncorrelated. Then, following the ISO *Guide* (Section 5), the associated standard uncertainty $u(x_{\rm C})$ is determined from the law of propagation of uncertainties: $u^2(x_{\rm C}) = \sum_i ({\rm d}X_{\rm C}/{\rm d}X_i)^2 u^2(x_i)$, where $\{{\rm d}X_{\rm C}/{\rm d}X_i\}$ are the partial derivatives of the function $X_{\rm C} = f(X_1, ..., X_n)$ with respect to $X_1, ..., X_n$ evaluated at $x_1, ..., x_n$, respectively. (When some of the laboratory values $X_1, ..., X_n$ are correlated, the expression for $u^2(x_{\rm C})$ should include covariance terms as discussed in the ISO *Guide*.) The combined result $x_{\rm C}$ is identified with the expected value $E(X_{\rm C})$ and the standard uncertainty $u(x_{\rm C})$ is identified with the standard deviation $S(X_{\rm C})$.

It is most convenient to choose the function $X_{\rm C} = f(X_1, ..., X_n)$ to be a linear function $X_{\rm C} = \sum_i a_i X_i$ for some specified weights $a_1, ..., a_n$, where $\sum_i a_i = 1$. In this case, the combined result $x_{\rm C} = \sum_i a_i x_i$ is equal to the expected value $E(X_{\rm C})$ and the squared standard uncertainty $u^2(x_{\rm C}) = \sum_i ({\rm d}X_{\rm C}/{\rm d}X_i)^2 u^2(x_i) = \sum_i a_i^2 u^2(x_i)$ is equal to the variance $V(X_{\rm C})$, which is the square of standard deviation $S(X_{\rm C})$. Two particular choices for the combined value $X_{\rm C}$ are the arithmetic mean $X_{\rm A}$ and the weighted mean $X_{\rm W}$ with weights $a_1, ..., a_n$ proportional to the reciprocals of the squared standard uncertainties. For the arithmetic mean $X_{\rm A} = (1/n)\sum_i X_i$, the combined result is $x_{\rm A} = (1/n)\sum_i x_i$ with squared standard uncertainty $u^2(x_{\rm A}) = (1/n^2)\sum_i u^2(x_i)$. (In the special case $u(x_1) = u(x_n) = u(x)$, the squared standard uncertainty (variance) $u^2(x_{\rm A})$ reduces to the familiar expression $u^2(x)/n$.) For the weighted mean $X_{\rm W} = \sum_i a_i X_i$ with weights $a_i = [1/u^2(x_i)]/\sum_i[1/u^2(x_i)]$, for $i = 1, 2, ..., n$, the combined result is $x_{\rm W} = \sum_i a_i x_i$ with squared standard uncertainty $u^2(x_{\rm W}) = \sum_i a_i^2 u^2(x_i) = 1/\sum_i[1/u^2(x_i)]$.

If some of the self-declared uncertainties $u(x_1), ..., u(x_n)$ are believed to be underestimated, a weighted mean $\sum_i a_i x_i$ with subjectively determined weights $a_1, ..., a_n$, where $\sum_i a_i = 1$, may be used as the uncorrected combined result $x_{\rm C}$. Sometimes, the experts involved in the interlaboratory evaluation choose to adjust the understated uncertainties. This amounts to using a weighted mean with subjectively determined weights $w_i = [1/\overline{u}^2(x_i)]/\sum_i[1/\overline{u}^2(x_i)]$, where $\overline{u}(x_1), ..., \overline{u}(x_n)$ are adjusted uncertainties, for $i = 1, 2, ..., n$.

When one or more of the results $x_1, ..., x_n$ are judged to be outliers, a weighted mean with subjectively determined weights may be used as the uncorrected combined result $x_{\rm C}$. Outliers are troublesome because they distort the arithmetic or weighted mean and the associated standard uncertainty. It is not easy to deal with outliers. A practical approach is to use a weighted mean that assigns zero weight to the outliers, then use them to determine the correction and uncertainty associated with the correction in Step 2 of the three-step method.

We do not recommend the weighted mean $x_{\rm W} = \sum_i a_i x_i$ with weights $a_i = [1/u^2(x_i)]/\sum_i[1/u^2(x_i)]$, for $i = 1, 2, ..., n$ as the reference value or the uncorrected combined result from those interlaboratory evaluations where not all self-declared uncertainties $u(x_1), ..., u(x_n)$ are reliable. When uncertainties $u(x_1), ..., u(x_n)$ are equal to the standard deviations of results $x_1, ..., x_n$ treated as random variables, the standard deviation (uncertainty) of the weighted mean $x_{\rm W}$ is smaller than that of the arithmetic mean $x_{\rm A}$. This is the motivation for using the weighted mean $x_{\rm W}$. In practice, uncertainties $u(x_1), ..., u(x_n)$ are estimates evaluated from statistical data (Type A) or scientific judgement (Type B). When some of these estimates are poor, the standard deviation (uncertainty) of the weighted mean $x_{\rm W}$ may actually exceed that of the arithmetic mean $x_{\rm A}$. A self-declared uncertainty may be poor because the uncertainty budget was not comprehensive. A Type A uncertainty may be unreliable because only a few independent measurements were made (see, for example, ISO *Guide*, Annex E.4). A Type B evaluation may be unreliable because the specified probability distribution under-represents or over-represents the uncertainty. Furthermore, if lacking thorough scientific and experimental knowledge, some laboratories may overstate or understate the uncertainties associated with their results.

We do not recommend the median of results $x_1, ..., x_n$ as the reference value or the uncorrected combined result from interlaboratory evaluations because it may not be statistically justified and is incompatible with the ISO *Guide*. Some (see, for example, [15]) suggest that when one or more of the results are judged to be discrepant, the median of results $x_1, ..., x_n$ should be used as the reference value because it is a robust statistic unaffected by a few discrepant results. When results $x_1, ..., x_n$ treated as random variables have the same sampling distribution, the median is indeed a robust statistic (see, for example, Rousseeuw [16]). However, the standard deviations of $x_1, ..., x_n$ may not be equal. That is, results $x_1, ..., x_n$ may not have the same distribution, so the median may not be justified. The expression of uncertainty associated with the median of results $x_1, ..., x_n$ is a multiple of the median of the absolute deviations of results $x_1, ..., x_n$ from their median. However, the standard uncertainty as defined by the ISO *Guide* is expressed as a standard deviation. So the median is incompatible with the *Guide*. Also, the standard uncertainty associated with the median of results $x_1, ..., x_n$ cannot be determined from the law of propagation of uncertainties. The only practical way to determine the standard uncertainty associated with the median of results $x_1$, ..., $x_n$ is numerical simulation of the median function $X_C = f(X_1, ..., X_n)$.

### *3.4 Corrected combined result* $y$ *and associated standard uncertainty* $u(y)$

According to Assumption II, a combined result $x_C$ (such as the arithmetic mean or a weighted mean) and its associated uncertainty $u(x_C)$ that place a non-negligible fraction of the results outside the 2-standard-uncertainty interval $[x_C \pm 2u(x_C)]$ are unsatisfactory representations of the information about $Y$ provided by $x_1$, ..., $x_n$. The inadequacy of such $x_C$ and $u(x_C)$ can be remedied by incorporating a correction for the difference between $x_C$ and $Y$ through the measurement equation: $Y = X_C + C = f(X_1, ..., X_n) + C$, where $X_1$, ..., $X_n$ are the laboratory values, $X_C = f(X_1, ..., X_n)$ is the uncorrected combined value, and $C$ is the correction, a random variable whose distribution represents belief about possible values of the difference $(Y - x_C)$, arising from systematic effects in results $x_1$, ..., $x_n$. According to Assumption II, the range of results $x_1$, ..., $x_n$ suggests possible values of $Y$ and hence of the difference $(Y - x_C)$. By definition, $C$ is independent of $X_C$. Let $E(C) = c$ and $S(C) = u(c)$ for some constants $c$ and $u(c)$. The random variable $C$ and its expected value $c$ are both referred to as a correction. The measurement equation $Y = X_C + C$ defines the corrected combined result for $Y$ as $y = x_C + c$ with standard uncertainty $u(y) = \sqrt{[u^2(x_C) + u^2(c)]}$. The result $y$ is identified with $E(Y)$ and the uncertainty $u(y)$ is identified with $S(Y)$. As the measurement equation $Y = X_C + C$ is linear, $y$ actually equals $E(Y)$ and $u(y)$ equals $S(Y)$.

### *3.5 Specification of correction* $E(C) = c$ *and associated standard uncertainty* $S(C) = u(c)$.

The proposed three-step approach to determine the corrected combined result $y$ and its associated standard uncertainty $u(y)$ is generic and permits the use of any probability distribution for correction $C$ that has finite expected value and finite standard deviation. As zero correction is reasonable and infinite correction is meaningless, we believe that the limits of the probability distribution for correction $C$ should be bounded and the interval between the limits should include zero. We use the symbols $-\alpha_1$ and $\alpha_2$ for the limits of the distribution of $C$, where $\alpha_1 \geq 0$ and $\alpha_2 \geq 0$. We discuss two simple distributions on the interval $(-\alpha_1, \alpha_2)$: a rectangular and a triangular distribution. The expected value, variance and standard deviation of a rectangular distribution on the interval $(-\alpha_1, \alpha_2)$ are $E(C) = (\alpha_2 - \alpha_1)/2$, $V(C) = (\alpha_2 + \alpha_1)^2/12$ and $S(C) = (\alpha_2 + \alpha_1)/\sqrt{12}$, respectively (see, for example, Evans et al. [14]). When $\alpha_1 = \alpha_2 = \alpha$, they reduce to $E(C) = 0$, and $V(C) = \alpha^2/3$, and $S(C) = \alpha/\sqrt{3}$, respectively. The expected value, variance and standard deviation of a triangular distribution on the interval $(-\alpha_1, \alpha_2)$ are $E(C) = (\alpha_2 - \alpha_1)/3$, $V(C) = (\alpha_2 - \alpha_1)^2/18 + (\alpha_1\,\alpha_2)/6$ and $S(C) = \sqrt{[(\alpha_2 - \alpha_1)^2/18 + (\alpha_1\,\alpha_2)/6]}$, respectively (see Appendix 3). When $\alpha_1 \neq \alpha_2$, the triangular distribution is asymmetric. Figure 2 shows the probability density function of an asymmetric triangular distribution. (Ayyangar [17] is credited for introducing asymmetric triangular distribution. However, we have parameterized it differently.) When $\alpha_1 = \alpha_2 = \alpha$, the expected value, variance and standard deviation of a triangular distribution reduce to $E(C) = 0$, $V(C) = \alpha^2/6$ and $S(C) = \alpha/\sqrt{6}$, respectively.

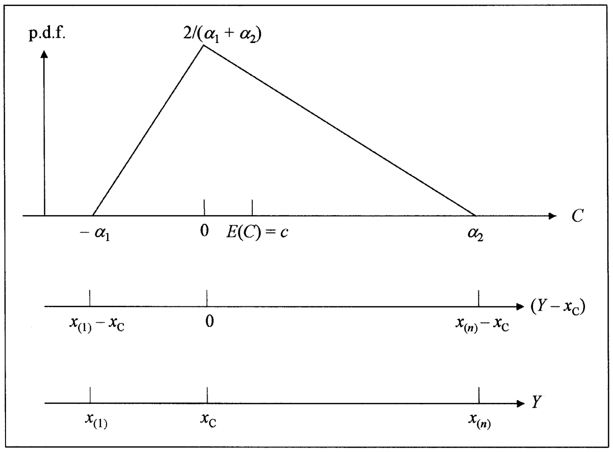

**Figure 2.** Probability density function (p.d.f.) of asymmetric triangular distribution $T(-\alpha_1, \alpha_2)$ in relation to the combined result $x_C$, the smallest result $x_{(1)} = \min\{x_1, x_2, ..., x_n\}$, and the largest result $x_{(n)} = \max\{x_1, x_2, ..., x_n\}$.

We suggest $-\alpha_1 = (x_{(1)} - x_C)$ and $\alpha_2 = (x_{(n)} - x_C)$, where $x_{(1)} = \min\{x_1, ..., x_n\}$ and $x_{(n)} = \max\{x_1, ..., x_n\}$ as default limits for the distribution of $C$. However, one may choose a wider pair of limits when plausible values of $Y$ are believed to exist outside the range of results $x_1$, ..., $x_n$. Such may be the case, for example, when $n$ is small.

Consider the following three common situations:

(a) All values of $Y$ in the range of results $x_1$, ..., $x_n$ are believed to be equally probable.

(b) The values of $Y$ near the middle of the range of results $x_1$, ..., $x_n$ are believed to be more probable than the values near the ends.

(c) The 2-standard-uncertainty interval $[y \pm 2u(y)]$ is required to include a specified range $(x_L, x_H)$ of results $x_1$, ..., $x_n$, where $x_L \leq x_C \leq x_H$. In particular, $x_L$ could be the minimum result $x_{(1)}$ and $x_H$ the maximum result $x_{(n)}$.

The adequacy of the correction $E(C) = c$ and the uncertainty $S(C) = u(c)$ is determined by three generic criteria:

(a) The 2-standard-uncertainty interval $[y \pm 2u(y)]$ should include a sufficiently large fraction of results $x_1, ..., x_n$.

(b) The absolute difference between the corrected and uncorrected combined results $|y - x_{\mathrm{C}}| = |c|$ should be small.

(c) The width of the 2-standard-uncertainty interval $[y \pm 2u(y)]$, which is affected by the size of $u(c)$, should not be too large.

When all values of $Y$ in the range of results $x_1, ..., x_n$ are believed to be equally probable, a rectangular distribution for $C$ with limits $-\alpha_1 = (x_{(1)} - x_{\mathrm{C}})$ and $\alpha_2 = (x_{(n)} - x_{\mathrm{C}})$ may be used, where $x_{(1)} = \min\{x_1, ..., x_n\}$ and $x_{(n)} = \max\{x_1, ..., x_n\}$. The correction $c$ and uncertainty $u(c)$ specified by such a rectangular distribution always yield an interval $[y \pm 2u(y)]$ that includes all results $x_1, ..., x_n$. This is a consequence of the property of rectangular distributions that the interval $[E(C) \pm 2S(C)] \equiv [c \pm 2u(c)]$ includes the entire range $(-\alpha_1, \alpha_2)$ of the distribution. However, the absolute correction $|c| = |\alpha_2 - \alpha_1|/2$ and the width of the 2-standard-uncertainty interval $[y \pm 2u(y)]$ determined from a rectangular distribution may be deemed to be too large. A rectangular distribution may not be unreasonable when $n$ is small.

When the values of $Y$ near the middle of the range of results $x_1, ..., x_n$ are believed to be more probable than the values near the ends, a triangular distribution for $C$ with limits $-\alpha_1 = (x_{(1)} - x_{\mathrm{C}})$ and $\alpha_2 = (x_{(n)} - x_{\mathrm{C}})$ may be used. The absolute correction $|c|$ and the width of the interval $[y \pm 2u(y)]$ determined from a triangular distribution are smaller than the corresponding specifications from rectangular distribution. The interval $[y \pm 2u(y)]$ determined from a triangular distribution may not include all results $x_1, ..., x_n$. However, the fraction of results $x_1, ..., x_n$ included by $[y \pm 2u(y)]$ may be sufficiently large. In terms of probability, a triangular distribution with limits $(x_{(1)} - x_{\mathrm{C}})$ and $\alpha_2 = (x_{(n)} - x_{\mathrm{C}})$ accommodates the viewpoint that the systematic effects in results $x_1, ..., x_n$ may be such that about half of the results are on each side of the value of the measurand. That is, the value of the measurand is more likely to fall near the middle rather than towards the ends of the range of results.

The shortest 2-standard-uncertainty interval $[y \pm 2u(y)]$ that includes a specified range $(x_{\mathrm{L}}, x_{\mathrm{H}})$ of results $x_1, ..., x_n$, can easily be determined when correction $c$ can be prescribed. For example, $c$ may be specified to be zero or some other small number. The choice of $c$ would depend on (a) the degree of asymmetry of the distribution of $C$ and (b) the percentage change in the result $x_{\mathrm{C}}$ that the chosen $c$ would imply. Thus, $y = x_{\mathrm{C}} + c$ is specified. Let $d = \max\{|x_{\mathrm{L}} - y|, |x_{\mathrm{H}} - y|\}$. Then the interval $[y \pm 2u(y)]$, where $u(y) = d/2$, is the shortest interval about $y$ that includes the specified range of results $(x_{\mathrm{L}}, x_{\mathrm{H}})$. The correction $c$ is what was specified and the uncertainty $u(c)$ is equal to $\surd[(d/2)^2 - u^2(x_{\mathrm{C}})]$, provided that $(d/2)^2 > u^2(x_{\mathrm{C}})$. In order to justify such $c$ and $u(c)$ as being technically consistent with the ISO *Guide*, we need to identify a probability distribution whose expected value is $c$ and whose standard deviation is $u(c)$. Appendix 4 gives the limits $-\alpha_1$ and $\alpha_2$ of rectangular and triangular distributions that yield a specified $c$ and $u(c)$. As the goal is to specify $c$, $u(c)$, $y$ and $u(y)$, the limits $-\alpha_1$ and $\alpha_2$ need not be calculated.

Thus a probability distribution for correction $C$ can be specified such that the 2-standard-uncertainty interval $[y \pm 2u(y)]$ would include a sufficiently large fraction of results $x_1, ..., x_n$, the absolute correction $|c|$ is small, and the width of the interval $[y \pm 2u(y)]$ is not too large.

Sometimes the nominal value $Y$ of the common measurand is specified in advance. In this case, the corrected combined result $y$ may be adjusted to match the nominal value $Y$. Such adjustment is made by adding or subtracting a constant from $y$ and $x_1, ..., x_n$. This does not affect the uncertainties associated with $y$ and $x_1, ..., x_n$.

## 4. Key comparison reference value and degree of equivalence

The KCRV, defined by the MRA [2], is the reference value (assigned to the common measurand) accompanied by its uncertainty resulting from a CIPM key comparison. According to the MRA, in most cases the KCRV can be considered to be a close, but not necessarily the best, approximation of the (corresponding) SI value. The most common choices for the KCRV are the arithmetic mean or a weighted mean of the results. We refer to them as the uncorrected combined result, denoted by $x_{\mathrm{C}}$. The associated standard uncertainty is denoted by $u(x_{\mathrm{C}})$. According to Assumption II, $x_{\mathrm{C}}$ and $u(x_{\mathrm{C}})$ are unsatisfactory representations of the information about $Y$ provided by results $x_1, ..., x_n$. Therefore, we suggest that the corrected combined result $y = x_{\mathrm{C}} + c$ and standard uncertainty $u(y) = \surd[u^2(x_{\mathrm{C}}) + u^2(c)]$ be used as the KCRV and its associated standard uncertainty, respectively.

The MRA [2] defines the degree of equivalence of a measurement standard as the degree to which the measurement standard (laboratory result) is consistent with the KCRV. This is expressed quantitatively by the deviation from the KCRV and the uncertainty of this deviation. Common choices are: (a) the pair $(x_i - x_{\mathrm{C}})$ and $u(x_i - x_{\mathrm{C}})$; (b) an uncertainty interval of the type $[(x_i - x_{\mathrm{C}}) \pm k\, u(x_i - x_{\mathrm{C}})]$ for some coverage factor $k$; and (c) the single value $(x_i - x_{\mathrm{C}})/u(x_i - x_{\mathrm{C}})$ for

$i = 1, 2, ..., n$, where $u(x_i - x_C)$ denotes the standard uncertainty associated with the deviation $(x_i - x_C)$ for $i = 1, 2, ..., n$. According to Assumption II, the KCRV should be $y$ with standard uncertainty $u(y)$. Therefore, we propose the expression $E_i = (x_i - y)/u(y)$ as the degree of equivalence between $x_i$ and $y$. This expression represents the deviation of $x_i$ from $y$ as a fraction of the standard uncertainty $u(y)$.

According to the MRA [2] the degree of equivalence between two measurement standards (two laboratory results) is the difference between their respective deviations from the KCRV and the uncertainty of this difference. Common choices are: (a) the pair $(x_i - x_j)$ and $u(x_i - x_j) = \surd[u^2(x_i) + u^2(x_j)]$; (b) an uncertainty interval of the type $[(x_i - x_j) \pm k\ \surd[u^2(x_i) + u^2(x_j)]]$ for some coverage factor $k$; and (c) the single value $(x_i - x_j)/\surd[u^2(x_i) + u^2(x_j)]$ for $i, j = 1, 2, ..., n$ and $i \neq j$. These expressions are based on only two results, $x_i$ and $x_j$, with uncertainties $u(x_i)$ and $u(x_j)$, respectively, for $i, j = 1, 2, ..., n$ and $i \neq j$. They ignore the information about $Y$ provided by other laboratories. If the authors of the MRA intended this, it is not clear why they defined the degree of equivalence as the difference between their respective deviations from the KCRV rather than the difference between the results. We propose the expression $E_i - E_j = [(x_i - y)/u(y) - (x_j - y)/u(y)]$ as the degree of equivalence between $x_i$ and $x_j$ for $i, j = 1, 2, ..., n$ and $i \neq j$.

## 5. International comparison of cryogenic radiometers

In order to illustrate the proposed three-step method, we have re-analysed some of the data from a recent BIPM Report on a supplementary comparison of cryogenic radiometers [15], initiated by the CCPR: CCPR-S3. Altogether seventeen national measurement institutes, including the BIPM, participated. All seventeen laboratories are believed to be competent. We briefly review the method adopted, which is described in detail in the BIPM Report, in order to identify the data that we have used for illustration.

The comparison of cryogenic radiometers was carried out indirectly by means of transfer standard detectors. The cryogenic radiometer at each laboratory was used to calibrate the responsivity of a set of three transfer standard detectors. When a set of detectors was received from a participating laboratory, it was compared with a set of continuously monitored control detectors at the BIPM before being shipped to the next laboratory. Any changes noticed in the transfer standard detectors and the effects of different experimental conditions in the participating laboratories were accounted for to bring the results reported by each laboratory to a common basis for comparison. The overall comparison was organized such that the independence of measurements from participating laboratories was maintained.

The relative difference between the responsivity of a transfer standard detector calibrated at a particular laboratory and the same detector calibrated at the BIPM was used as the basis for comparison of cryogenic radiometers. Each laboratory used at least three out of six possible wavelengths for laser sources. The argon line with a wavelength of 514.536 nm was used by all. The BIPM Report [15, Section 1.3] denotes the responsivity of a detector calibrated at laboratory A as $R_A$ and the responsivity of the same detector calibrated at the BIPM as $R_{BIPM}$. The relative difference $\Delta$ is defined as $\Delta = (R_A - R_{BIPM})/R_{BIPM}$. The relative difference for the three detectors in a set is averaged. The corresponding uncertainty, denoted by $u_C$ in the BIPM Report, combines the relative standard uncertainties from each laboratory and the uncertainties associated with the transfer standard detectors. The BIPM Report (Table 65, columns 6 and 7) displays data on $\Delta \times 10^4$ and $u_C \times 10^4$ for wavelength 514.536 nm. These data are reproduced here as Table 1. We have re-analysed these data. In particular, we treat the data on the relative differences from the BIPM measurement as laboratory results $x_1, x_2, ..., x_{17}$ with associated standard uncertainties $u(x_1), u(x_2), ..., u(x_{17})$. The relative difference $x_{17}$, for the BIPM laboratory, is defined to be zero. The associated uncertainty $u(x_{17}) = 1.0 \times 10^{-4}$ represents the uncertainty from the BIPM uncertainty budget [15, Tables 5 and 6]. This uncertainty component is also included in the other sixteen results.

The Birge ratio for the data in Table 1 is $R_B = 1.21$. (This is not significantly larger than 1; the probability of a chi-square distribution with $(n - 1) = 16$ degrees of freedom exceeding $(n - 1)$

**Table 1.** Relative differences from the BIPM measurement for wavelength 514.536 nm and their associated standard uncertainties, reproduced from the BIPM Report [15], Table 65, columns 6 and 7.

| Laboratory names and indices | | Relative difference $x_i \times 10^4$ | Standard uncertainty $u(x_i) \times 10^4$ |
|---|---|---|---|
| 1 | PTB.T | –0.20 | 1.30 |
| 2 | BNM.INM | 1.10 | 1.70 |
| 3 | CSIRO | 2.00 | 1.40 |
| 4 | DFM | –0.30 | 2.50 |
| 5 | ETL | 13.10 | 4.90 |
| 6 | HUT | 1.70 | 2.70 |
| 7 | IEN | –11.00 | 6.80 |
| 8 | IFA | 0.00 | 2.20 |
| 9 | MSL | 0.30 | 1.30 |
| 10 | KRISS | –5.10 | 2.40 |
| 11 | NIST | 5.90 | 3.20 |
| 12 | NMI-VSL | –1.10 | 2.60 |
| 13 | NPL | 1.30 | 1.10 |
| 14 | NRC | 5.30 | 3.40 |
| 15 | PTB.R | 2.90 | 2.90 |
| 16 | SP | –1.00 | 5.10 |
| 17 | BIPM | 0.00 | 1.00 |

**Table 2.** The arithmetic mean $x_\mathrm{A}$ as the uncorrected combined result, uncertainty $u(x_\mathrm{A})$; correction $c$, uncertainty $u(c)$; and corrected combined result $y$, uncertainty $u(y)$. The formulae used for these computations are given in Appendix 5.

| Component | Result | Standard uncertainty |
|---|---|---|
| Uncorrected combined result | $x_\mathrm{A} \times 10^4 = 0.88$ | $u(x_\mathrm{A}) \times 10^4 = 0.76$ |
| Correction | $c \times 10^4 = 0.12$ | $u(c) \times 10^4 = 4.92$ |
| Corrected combined result | $y \times 10^4 = 0.99$ | $u(y) \times 10^4 = 4.98$ |

times $R_\mathrm{B}^2$ is 10 %.) If we remove the two extreme results corresponding to laboratories 5 and 7, the Birge ratio drops to $R_\mathrm{B} = 1.00$. So the results given in Table 1 are consistent according to the Birge ratio test. That is, no results are discrepant. In particular, results $x_5$ and $x_7$ are not discrepant.

The BIPM Report investigated the arithmetic mean, the weighted mean with weights proportional to the reciprocals of the squared standard uncertainties $u(x_1)$, $u(x_2)$, ..., $u(x_{17})$, and the median of results $x_1$, $x_2$, ..., $x_{17}$ for common reference. Finally, the BIPM Report chose the weighted mean as the CCPR reference value. What does the CCPR reference value represent? Indeed, what is the common measurand of value $Y$ whose estimates are the results $x_1$, $x_2$, ..., $x_{17}$ with their respective uncertainties $u(x_1)$, $u(x_2)$, ..., $u(x_{17})$? Given results $x_1$, $x_2$, ..., $x_{17}$, which are relative differences from the BIPM measurement, we suggest *modus operandi* that value $Y$ may be defined as the relative difference from the BIPM measurement that might be realized by *any* competent laboratory. One of the objects of the interlaboratory evaluation is to quantify the worldwide uncertainty in cryogenic radiometric measurements by competent laboratories. The standard uncertainty associated with $Y$ is a quantitative measure of the worldwide uncertainty provided by results $x_1$, $x_2$, ..., $x_{17}$ and uncertainties $u(x_1)$, $u(x_2)$, ..., $u(x_{17})$.

Let $X_1$, $X_2$, ..., $X_{17}$ represent the relative differences for the laboratories that participated in the international comparison. The random variables $X_1$, $X_2$, ..., $X_{17}$ represent the states of knowledge about $Y$ within the individual laboratories. Following the ISO *Guide*, laboratory results $x_1$, $x_2$, ..., $x_{17}$ are identified with the expected values of $X_1$, $X_2$, ..., $X_{17}$, respectively. The uncertainties $u(x_1)$, $u(x_2)$, ..., $u(x_{17})$ are identified with the standard deviations of $X_1$, $X_2$, ..., $X_{17}$, respectively. For illustration, we have used the arithmetic mean $X_\mathrm{A} = (1/n)\sum_i X_i$ as the uncorrected combined value. The corresponding combined result is $x_\mathrm{A} = (1/n)\sum_i x_i$ and the associated standard uncertainty is $u(x_\mathrm{A}) = (1/n)\ \surd[\sum_i u^2(x_i)]$, where $n = 17$. Table 2 gives the computed values of $x_\mathrm{A}$ and $u(x_\mathrm{A})$. Figure 1 plots the laboratory results $x_1$, $x_2$, ..., $x_{17}$, the uncorrected combined result (u.c.r.) $x_\mathrm{A}$, and the associated 2-standard-uncertainty intervals. We note that the 2-standard-uncertainty interval $[x_\mathrm{A} \pm 2u(x_\mathrm{A})]$ associated with the arithmetic mean $x_\mathrm{A}$ excludes five, i.e. 29 %, of the seventeen results $x_1$, ..., $x_{17}$. Also, several other results are on the borderline. As the excluded results are from competent laboratories, the result $x_\mathrm{A}$ and the uncertainty $u(x_\mathrm{A})$ are unsatisfactory representations of the relative difference $Y$ from the BIPM measurement that might be realized by any competent laboratory. The inadequacy of $x_\mathrm{A}$ and $u(x_\mathrm{A})$ can be remedied by incorporating a correction and the uncertainty associated with it for possible difference between $x_\mathrm{A}$ and $Y$. For illustration, we have chosen an asymmetric triangular distribution for the correction denoted by $C$ with limits $-\alpha_1 = (x_{(1)} - x_\mathrm{A})$ and $\alpha_2 = (x_{(n)} - x_\mathrm{A})$. This distribution represents the belief that a competent laboratory is more likely to realize a relative difference from the BIPM measurement in the vicinity of $x_\mathrm{A}$ than far from $x_\mathrm{A}$. The corresponding expected value $E(C) = c$ and the standard uncertainty $S(C) = u(c)$ are given in Table 2. The formulae used for computing $c$ and $u(c)$ are given in Appendix 5. The corrected combined result $y = x_\mathrm{A} + c$ and the standard uncertainty $u(y) = \surd[u^2(x_\mathrm{A}) + u^2(c)]$ are also given in Table 2. The result $y$ and the uncertainty $u(y)$ represent the distribution of $Y$. Figure 3 plots the laboratory results $x_1, x_2$, ..., $x_{17}$ and the corrected combined result (c.c.r.) $y$, with associated 2-standard-uncertainty intervals. The centre line is drawn at the corrected combined result $y$ and the dashed lines are drawn at $y - 2u(y)$ and $y + 2u(y)$, respectively. We note that the 2-standard-uncertainty interval $[y - 2u(y),\ y + 2u(y)]$ associated with the corrected combined result $y$ includes all but

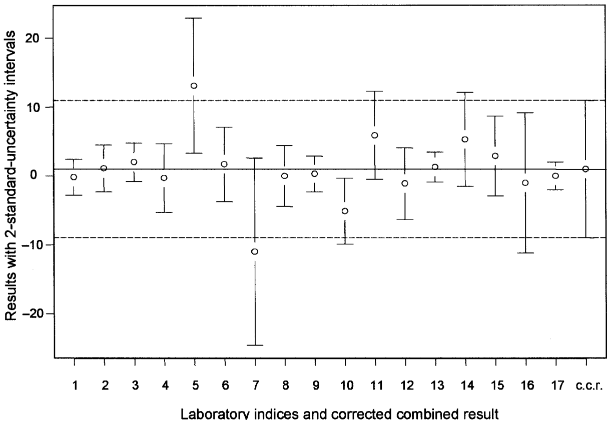

**Figure 3.** Individual laboratory results and the corrected combined result $y = x_\mathrm{A} + c$ with associated 2-standard-uncertainty intervals. Table 1 lists the laboratories corresponding to indices 1, ..., 17 and c.c.r. stands for corrected combined result. The horizontal centre line is drawn at the corrected combined result $y$. The dashed lines are drawn at the limits $y - 2u(y)$ and $y + 2u(y)$ of the 2-standard-uncertainty interval associated with $y$.

the two extreme results. The 2-standard-uncertainty intervals associated with the two excluded results $x_5$ and $x_7$ have substantial intersection with the interval $[y - 2u(y),\ y + 2u(y)]$. So the corrected combined result $y$ and uncertainty $u(y)$ represent better than $x_\mathrm{A}$ and $u(x_\mathrm{A})$ the relative difference $Y$ from the BIPM measurement that might be realized by any competent laboratory. That is, $y$ and $u(y)$ represent the worldwide uncertainty better than $x_\mathrm{A}$ and $u(x_\mathrm{A})$.

It is not unreasonable to set the CCPR reference value as zero in this example. This can be done by subtracting the corrected combined result $y$ from each of the results $x_1$, $x_2$, ..., $x_{17}$. In particular, result $x_{17}$ from the BIPM would be adjusted to $-y$. The uncertainty $u(y)$ and the degree of equivalence $E_i = (x_i - y)/u(y)$ for $i = 1$, 2, ..., $n$ are unaffected by such adjustment.

If we had used a rectangular distribution for correction $C$ with limits $-\alpha_1 = (x_{(1)} - x_\mathrm{A})$ and $\alpha_2 = (x_{(n)} - x_\mathrm{A})$, we would have $c = 0.17 \times 10^{-4}$, $u(c) = 6.96 \times 10^{-4}$, $y = 1.05 \times 10^{-4}$, and $u(y) = 7.00 \times 10^{-4}$. If we had used a symmetric distribution for correction $C$ with the requirement that the interval $[y \pm 2u(y)]$ that includes all seventeen results be the shortest, we would have $d = \max\{|x_{(1)} - x_\mathrm{A}|, |x_{(n)} - x_\mathrm{A}|\} = 12.22 \times 10^{-4}$, $y = 0.88 \times 10^{-4}$, and $u(y) = 6.11 \times 10^{-4}$.

We noticed certain technical deficiencies in the analysis of the BIPM Report. They are discussed in Appendix 6. These deficiencies point to the need for a generally accepted approach for the data analysis of interlaboratory evaluations, including CIPM key comparisons.

## 6. Summary

We address the problem of determining the combined result and its associated uncertainty in the measurement of a common measurand by a group of competent laboratories. Most data analyses of interlaboratory evaluations are based on the highly questionable assumption that the expected values of the individual laboratory results $x_1$, ..., $x_n$ are all equal to the value $Y$ of the common measurand. This means that the laboratory results are subject to random effects only with respect to the value of the measurand. We use the more realistic assumption that results $x_1$, ..., $x_n$ are subject to both random and systematic effects with respect to the value of the measurand. The value $Y$ of the measurand may be anywhere in the range of results $x_1$, ..., $x_n$, or even outside this range when $n$ is small. Therefore, a combined result and its associated standard uncertainty that place a non-negligible fraction of the results outside the 2-standard-uncertainty interval are unsatisfactory representations of the information about $Y$ provided by the set of results $x_1$, ..., $x_n$ and their associated uncertainties.

Previous efforts to account for the uncertainty arising from systematic effects are either not consistent with the ISO *Guide* or of limited applicability. We address the general case where the number $n$ of laboratories is arbitrary and the combined result may be based on the arithmetic mean or a weighted mean. Following the approach of the ISO *Guide* in dealing with systematic effects, we propose a three-step method to determine a combined result $y$ and its associated uncertainty $u(y)$.

*Step 1*: *Determine the uncorrected combined result $x_\mathrm{C}$ and its associated standard uncertainty denoted by $u(x_\mathrm{C})$. Assess the need for correcting the result $x_\mathrm{C}$.* We suggest that the arithmetic mean should be used as the default uncorrected combined result $x_\mathrm{C}$. A subjectively determined weighted mean may be used when justified. Such would be the case, for example, when one or more of the results are outliers or some of the self-declared uncertainties associated with the individual results are believed to be understatements and the experts choose to adjust them. Determine the standard uncertainty $u(x_\mathrm{C})$ from the uncertainties $u(x_1)$, ..., $u(x_n)$ associated with results $x_1$, ..., $x_n$ using the law of propagation of uncertainties. Assess the need for correcting result $x_\mathrm{C}$ for a possible difference between $x_\mathrm{C}$ and the value $Y$ of the measurand. Such a correction is needed whenever the interval $[x_\mathrm{C} \pm 2u(x_\mathrm{C})]$ excludes a non-negligible fraction of the results $x_1$, ..., $x_n$.

*Step 2*: *Determine the correction $E(C) = c$ to be applied to $x_\mathrm{C}$ and the standard uncertainty $S(C) = u(c)$ associated with the correction.* Following the ISO *Guide*, a probability distribution is used to specify correction $c$ and its associated uncertainty $u(c)$. The probability distribution represents belief about reasonable correction. When all values of $Y$ in the range of results $x_1$, ..., $x_n$ are believed to be equally probable, a rectangular distribution for $C$ may be used. When the values of $Y$ near the middle of the range of results $x_1$, ..., $x_n$ are believed to be more probable than the values near the ends, a triangular distribution for $C$ may be used. The limits of the probability distribution of $C$ can be determined from the range of deviations $(x_1 - x_\mathrm{C})$, $(x_2 - x_\mathrm{C})$, ..., $(x_n - x_\mathrm{C})$ of the results $x_1$, $x_2$, ..., $x_n$ from the uncorrected combined result $x_\mathrm{C}$.

*Step 3*: *Determine the corrected combined result $y = x_\mathrm{C} + c$ and the associated combined standard uncertainty $u(y) = \surd[u^2(x_\mathrm{C}) + u^2(c)]$.* With a judiciously specified probability distribution for correction $C$, the 2-standard-uncertainty interval $[y \pm 2u(y)]$ would include a sufficiently large fraction of results $x_1$, ..., $x_n$. Thus $y$ and $u(y)$ would represent, better than $x_\mathrm{C}$ and $u(x_\mathrm{C})$, the information about $Y$ provided by results $x_1$, ..., $x_n$.

When the interlaboratory evaluation is a CIPM key comparison, we suggest that the combined result $y$ and its associated standard uncertainty $u(y)$ determined by the three-step method be identified with the key comparison reference value and its associated uncertainty. These two quantities can then be used to specify the degree of equivalence of the individual results.

**Acknowledgements.** We are grateful to R. Goebel, R. Köhler and M. Stock of the BIPM for clarifying the details of their analysis. We thank Geoffrey McFadden and Ron Boisvert for their support. Discussions with Peter Mohr and Barry Taylor were very useful. The following provided useful comments on an earlier draft: Dipak Dey, Yoshi Ohno, Ingram Olkin, Tom Vetter, and the referee.

## Appendix 1

### *Terminology of the ISO* Guide

*Measurand*: particular quantity subject to measurement (ISO *Guide*, B.2.9). Denoted by $Y$.

*Result of a measurement*: value attributed to a measurand, obtained by measurement (ISO *Guide*, B.2.11). Denoted by $y$.

*Uncertainty of measurement*: parameter, associated with the result of a measurement, that characterizes the dispersion of the values that could reasonably be attributed to the measurand (ISO *Guide*, B.2.18).

*Standard uncertainty*: uncertainty of the result of a measurement expressed as a standard deviation (ISO *Guide*, 2.3.1). Denoted by $u(y)$.

*Combined standard uncertainty*: determined by combining the individual standard uncertainties (and covariances as appropriate) using the root-sum-of-squares method, or equivalent established and documented methods (based on NIST TN-1297, C.2.2).

*Expanded uncertainty*: multiple of combined standard uncertainty; the multiplier is called *coverage factor* (based on NIST TN-1297, C.2.3). Denoted by $U = ku(y)$. The conventional value of the coverage factor is $k = 2$. The corresponding expanded uncertainty is referred to as 2-standard uncertainty.

*Expanded uncertainty interval*: interval defined by expanded uncertainty about the result of a measurement that may be expected to encompass a large fraction of the distribution of values that could reasonably be attributed to the measurand (based on ISO *Guide*, 2.3.5). Denoted by $[y \pm ku(y)]$. The expanded uncertainty interval with coverage factor $k = 2$, i.e. $[y \pm 2u(y)]$ is referred to as 2-standard-uncertainty interval.

*Measurement equation*: equation that represents the value of the measurand as a function of all those quantities that contribute to the determination of the corresponding result of measurement and the associated combined standard uncertainty. All quantities involved in a measurement equation are treated as random variables with finite expected values, variances, and covariances (as appropriate). Each input quantity of a measurement equation may have its own measurement equation (based on the website http://physics.nist.gov/cuu/Uncertainty/basic.html, and on NIST TN-1297, D.3.1).

## Appendix 2

### *Birge ratio as a statistical estimate*

The Birge ratio can be described as a statistical estimate from the Aitken [18] weighted-least-squares model. The Aitken weighted-least-squares model is $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{e}$, where $E(\boldsymbol{e}) = \boldsymbol{0}$ and $V(\boldsymbol{e}) = \sigma^2\boldsymbol{V}$. Here, $\boldsymbol{y}$ is an $n \times 1$ vector of random measurements, $\boldsymbol{X}$ is an $n \times p$ known matrix of rank $p$, $\boldsymbol{\beta}$ is an $n \times 1$ vector of unknown parameters, $\boldsymbol{e}$ is an $n \times 1$ vector of random errors, $\sigma^2$ is an unknown parameter, and $\boldsymbol{V}$ is an $n \times n$ known positive definite matrix. The "best linear unbiased estimate" of $\boldsymbol{\beta}$ is $\boldsymbol{b} = (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{y}$ with variance $V(\boldsymbol{b}) = \sigma^2\,(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}$. The analysis of variance (ANOVA) estimate of $\sigma^2$ is $s^2 = (\boldsymbol{y} - \boldsymbol{Xb})'\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{Xb})/(n-p)$ and $E(s^2) = \sigma^2$. When the distribution of $\boldsymbol{y}$ can be assumed to be multivariate normal, $(n-p)\,s^2/\sigma^2$ has a chi-square distribution with $(n-p)$ degrees of freedom (see, e.g. Rao [19]).

The Birge ratio model may be written as $x_i = \mu + e_i$, where $e_1, ..., e_n$ are independent random errors with $E(e_i) = 0$ and $V(e_i) = u^2(x_i)$ for $i = 1, ..., n$. This model is a special case of the Aitken weighted-least-squares model, where $\boldsymbol{y}$ is the vector of laboratory results $(x_1, ..., x_n)'$, $\boldsymbol{X}$ is a vector of ones $(1, ..., 1)'$, $\boldsymbol{\beta}$ is $\mu$, $\sigma^2 \equiv 1$, and $\boldsymbol{V}$ is the known diagonal matrix, $\mathrm{diag}[u^2(x_1), ..., u^2(x_n)]$. By making these substitutions, the "best linear unbiased estimate" of $\mu$ is $x_{\mathrm{W}} = \sum_i w_i x_i / \sum_i w_i$ with variance $V(x_{\mathrm{W}}) = \sigma^2/\sum_i w_i$, where $w_i = 1/u^2(x_i)$ for $i = 1, ..., n$ and $\sigma^2 \equiv 1$. The ANOVA estimate of $\sigma^2 \equiv 1$ is $s^2 = \sum_i w_i(x_i - x_{\mathrm{W}})^2/(n-1)$. Thus the Birge ratio $R_{\mathrm{B}} = \sqrt{[\sum_i w_i(x_i - x_{\mathrm{W}})^2/(n-1)]}$ is the ANOVA estimate of $\sigma \equiv 1$ in Aitken weighted-least-squares model and $E(R_{\mathrm{B}}^2) = 1$. When results $x_1, ..., x_n$ can be assumed to be normally distributed, $(n-1)\,R_{\mathrm{B}}^2$ has a chi-square distribution with $(n-1)$ degrees of freedom. In this case the probability due to random chance of realizing the Birge ratio as large as observed or larger can be quantified. Birge [3] called $V(x_{\mathrm{W}}) = 1/\sum_i w_i$ a measure of "internal consistency" denoted by $\sigma_{\mathrm{I}}^2$ and called $[\sum_i w_i(x_i - x_{\mathrm{W}})^2/(n-1)] \times [1/\sum_i w_i]$ a measure of "external consistency" denoted by $\sigma_{\mathrm{E}}^2$. He defined $R_{\mathrm{B}}^2 = \sigma_{\mathrm{E}}^2/\sigma_{\mathrm{I}}^2$.

## Appendix 3

### *Expected value and standard deviation of triangular distribution*

The probability density function of a random variable $T$ having a triangular distribution on the interval

$(-\alpha_1, \alpha_2)$ is as follows: $p(t) = [2/(\alpha_1 + \alpha_2)] \times [1 + t/\alpha_1]$, when $-\alpha_1 \le t \le 0$; $p(t) = [2/(\alpha_1 + \alpha_2)] \times [1 - t/\alpha_2]$, when $0 \le t \le \alpha_2$; and $p(t) = 0$, otherwise. Figure 2 shows this probability density function (p.d.f.). It can be checked that $\int p(t)\mathrm{d}t = 1$, which is the same as saying that the area of the triangle is one. Simple integration shows that expected value $E(T) = \int t\, p(t)\mathrm{d}t = (\alpha_2 - \alpha_1)/3$ and $E(T^2) = \int t^2\, p(t)\mathrm{d}t = [(\alpha_2 - \alpha_1)^2/6] + [\alpha_1\alpha_2/6]$. Thus, variance $V(T) = (\alpha_2 - \alpha_1)^2/18 + (\alpha_1\alpha_2)/6$ and standard deviation $S(T) = \surd[V(T)]$. Two equivalent formulae for $V(T)$ are $E(T)^2/2 + (\alpha_1\alpha_2)/6$ and $(\alpha_1^2 + \alpha_2^2 + \alpha_1\alpha_2)/18$.

## Appendix 4

### *Limits of rectangular and triangular distributions for given $c$ and $u(c)$*

For a rectangular distribution with limits $-\alpha_1$ and $\alpha_2$, $E(C) = c = (\alpha_2 - \alpha_1)/2$, and $V(C) = u^2(c) = (\alpha_2 + \alpha_1)^2/12$. By solving for $\alpha_1$ and $\alpha_2$, we get $\alpha_1 = \surd 3u(c) - c$ and $\alpha_2 = \surd 3u(c) + c$. For a triangular distribution with limits $-\alpha_1$ and $\alpha_2$, $E(C) = c = (\alpha_2 - \alpha_1)/3$, and $V(C) = u^2(c) = (\alpha_2 - \alpha_1)^2/18 + (\alpha_1\alpha_2)/6$. Therefore, $(\alpha_2 + \alpha_1)^2 = (\alpha_2 - \alpha_1)^2 + 4\alpha_1\alpha_2 = 24u^2(c) - 3c^2$. By solving for $\alpha_1$ and $\alpha_2$, we get $2\alpha_1 = \surd[24u^2(c) - 3c^2] - 3c$ and $2\alpha_2 = \surd[24u^2(c) - 3c^2] + 3c$.

## Appendix 5

### *Formulae used for computations given in Table 2*

Uncorrected combined result: $x_\mathrm{A} = (1/n)\sum_i x_i$.
Standard uncertainty: $u(x_\mathrm{A}) = (1/n)\surd[\sum_i u^2(x_i)]$.
Limits of the triangular distribution for correction $C$: $-\alpha_1 = (x_{(1)} - x_\mathrm{A})$ and $\alpha_2 = (x_{(n)} - x_\mathrm{A})$, where $x_{(1)} = \min\{x_1, ..., x_n\}$ and $x_{(n)} = \max\{x_1, ..., x_n\}$.
Correction: $E(C) = c = (\alpha_2 - \alpha_1)/3$.
Standard uncertainty: $S(C) = u(c) = \surd[(\alpha_2 - \alpha_1)^2/18 + (\alpha_1\alpha_2)/6]$.
Corrected combined result: $y = x_\mathrm{A} + c$.
Standard uncertainty: $u(y) = \surd[u^2(x_\mathrm{A}) + u^2(c)]$.

## Appendix 6

### *Discussion of BIPM Report*

In the BIPM Report [15], the uncertainties $u(x_1), ..., u(x_n)$ are computed as *relative standard uncertainties* but used as *absolute standard uncertainties*. In our calculations, we have also used the uncertainties $u(x_1), ..., u(x_n)$ as absolute standard uncertainties. Our calculations show that the weighted mean is $x_\mathrm{W} = 0.65 \times 10^{-4}$ and the absolute standard uncertainty is $u(x_\mathrm{W}) = 0.44 \times 10^{-4}$. We used the formulae $x_\mathrm{W} = \Sigma a_i x_i$, where $a_i = [1/u^2(x_i)]/\sum_i[1/u^2(x_i)]$, for $i = 1, 2, ..., n$, and $u^2(x_\mathrm{W}) = 1/\sum_i[1/u^2(x_i)]$. Our calculations agree with the corresponding results $x_\mathrm{W} = 0.7 \times 10^{-4}$ and $u(x_\mathrm{W}) = 0.4 \times 10^{-4}$ given in the BIPM Report (Tables 63 and 64). We can, therefore, conclude that the BIPM Report used the formulae given in its Sections 5.2.1 and 5.2.2. In these formulae the uncertainties $u(x_1), ..., u(x_n)$ are absolute standard uncertainties. However, Section 4 of the BIPM Report computes the uncertainties $u(x_1), ..., u(x_n)$ as relative standard uncertainties.

The formula used in the BIPM Report for the uncertainty associated with the median of results $x_1, ..., x_n$ is unsatisfactory. The BIPM Report evaluated the uncertainty associated with the median as $1.9/\surd(n - 1)$ times the median of the absolute deviations of results $x_1, ..., x_n$ from their median. This formula does not agree with the ISO *Guide*, which expresses all uncertainties as standard uncertainties or their multiples. Also, it assumes that the standard deviations of results $x_1, ..., x_n$ are equal. The BIPM Report does not justify this assumption. The uncertainties $u(x_1), ..., u(x_n)$ range from $1.00 \times 10^{-4}$ to $6.80 \times 10^{-4}$. (The calculated value of the test statistic $\max\{u^2(x_1), ..., u^2(x_n)\}/\sum_i u^2(x_i)$ is 0.28. This is not insignificant, see Pearson and Hartley [20]).

The formula used in the BIPM Report for the standard uncertainty $u(x_\mathrm{A})$ associated with the arithmetic mean $x_\mathrm{A}$ is unsatisfactory. Our result $x_\mathrm{A} = 0.88 \times 10^{-4}$ agrees with the result $x_\mathrm{A} = 0.9 \times 10^{-4}$ given in the BIPM Report (Table 63). But our result $u(x_\mathrm{A}) = 0.76 \times 10^{-4}$ does not agree with the result $1.2 \times 10^{-4}$ given in the BIPM Report (Table 64). The BIPM Report evaluated the standard uncertainty associated with the arithmetic mean as $s_x/\surd n = \surd[\sum_i(x_i - x_\mathrm{A})^2]/\surd[n(n - 1)]$, where $x_1, ..., x_n$ are laboratory results. The use of $s_x/\surd n$ as the standard uncertainty $u(x_\mathrm{A})$ associated with the arithmetic mean $x_\mathrm{A}$ does not parallel the approach used by the BIPM Report to evaluate the standard uncertainty $u(x_\mathrm{W})$ associated with the weighted mean $x_\mathrm{W}$. For $u(x_\mathrm{W})$, the BIPM Report used the formula $u(x_\mathrm{W}) = \surd[\sum_i a_i^2 u^2(x_i)]$, where $a_i = [1/u^2(x_i)]/\sum_i[1/u^2(x_i)]$ for $i = 1, 2, ..., n$. This formula reduces to $u(x_\mathrm{W}) = \surd\{1/\sum_i[1/u^2(x_i)]\}$ as stated in Sections 5.2.1 and 5.2.2 of the BIPM Report. The corresponding formula for $u(x_\mathrm{A})$ is $u(x_\mathrm{A}) = \surd[\sum_i a_i^2 u^2(x_i)]$, where $a_i = 1/n$ for $i = 1, 2, ..., n$. This formula reduces to $u(x_\mathrm{A}) = (1/n)\surd[\sum_i u^2(x_i)]$. For the data in Table 1, $u(x_\mathrm{A}) = (1/n)\surd[\sum_i u^2(x_i)] = 0.76 \times 10^{-4}$.

## References

1. *Guide to the Expression of Uncertainty in Measurement*, 2nd ed., Geneva, International Organization for Standardization, 1995, ISBN 92-67-10188-9. (The following seven organizations supported the development of this

*Guide*, which is published in their names. International Bureau of Weights and Measures (BIPM), International Electrotechnical Commission (IEC), International Federation of Clinical Chemistry (IFCC), International Organization for Standardization (ISO), International Union of Pure and Applied Chemistry (IUPAC), International Union of Pure and Applied Physics (IUPAP), and International Organization of Legal Metrology (OIML). The American National Standards Institute (ANSI) and the NCSL International have adopted the ISO *Guide* as the *U.S. Guide to the Expression of Uncertainty in Measurement*, ANSI/NCSL Z540-2-1997, available from NCSL International, 1800, 30th Street, Suite 305B, Boulder, CO 80301, USA.)

2. Mutual Recognition Arrangement, 28 February 2001, (http://www.bipm.org/enus/8_Key_Comparisons/key_comparisons.html).
3. Birge R. T., *Phys. Rev.*, 1932, **40**, 207-227.
4. Cochran W. G., *J. R. Stat. Soc.*, 1937, **4** (Suppl.), 102-118.
5. Schiller S. B., Eberhardt K. R., *Spectrochimica Acta*, 1991, **46B**, 1607-1613.
6. Levenson M. S., Banks D. L., Eberhardt K. R., Gill L. M., Guthrie W. F., Liu H. K., Vangel M. G., Yen J. H., Zhang N. F., *J. Res. Natl. Inst. Stand. Technol.*, 2000, **105**, 571-579.
7. Paule R. C., Mandel J., *J. Res. Natl. Bur. Stand.*, 1982, **87**, 377-385.
8. Rukhin A. L., Vangel M. G., *J. Am. Stat. Assoc.*, 1998, **93**, 303-308.
9. Taylor B. N., Kuyatt C. E., *Guidelines for Evaluating and Expressing the Uncertainty of NIST Measurement Results*, NIST Technical Note 1297, U.S. Department of Commerce, 1994. (This report elaborates on the ISO *Guide* and includes the NIST policy on expression of uncertainty.)
10. Taylor B. N., Parker W. H., Langenberg D. N., *Rev. Mod. Phys.*, 1969, **41**, 375-496 (Section III.A.1.).
11. Mohr P. J., Taylor B. N., *J. Phys. Chem. Ref. Data*, 1999, **28**, 1713-1852, Section 4.1 and Appendix E.
12. Hoel P. G., *Introduction to Mathematical Statistics*, 4th ed., New York, John Wiley & Sons, 1971.
13. Box G. E. P., Tiao G. C., *Bayesian Inference in Statistical Analysis*, Theorem 2.4.1, Reading, Mass., Addison-Wesley, 1973.
14. Evans M., Hastings N., Peacock B., *Statistical Distributions*, 3rd ed., New York, John Wiley & Sons, 2000.
15. Goebel R., Stock M., Köhler R., Report on the International Comparison of Cryogenic Radiometers Based on Transfer Detectors, *BIPM Rapport BIPM-2000/9*, Sèvres, Bureau International des Poids et Mesures, September 2000 (http://www.bipm.org/pdf/RapportBIPM/2000/09.pdf).
16. Rousseeuw P. J., Robust Estimation and Identifying Outliers, In: *Handbook of Statistical Methods for Engineers and Scientists* (Edited by Harrison M. Wadsworth), 2nd ed., Chapter 17, New York, McGraw Hill, 1998.
17. Ayyangar A. A. K., The Triangular Distribution, In: *Mathematics Student*, Vol. 9, 85-87, 1941. (Discussed in N. L. Johnson, S. Kotz and N. Balakrishnan, *Continuous Univariate Distributions*, Vol. 2, 2nd ed., Chapter 26, Section 9, New York, John Wiley & Sons, 1995.)
18. Aitken A. C., Proc. *R. Soc. Edinburgh, A*, 1935, **55**, 42-48.
19. Rao C. R., *Linear Statistical Inference and its Applications*, 2nd ed., Section 4a, New York, John Wiley & Sons, 1973.
20. Pearson E. S., Hartley H. O., *Biometrika Tables for Statistician*, 3rd ed., Table 31a, Cambridge, Cambridge University Press, 1966.

*Received on 19 December 2001 and in revised form on 11 April 2002.*